FEB 17 2004

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): Paul St. John BRITTAN, et al.) Group: 2641

) Examiner: Not yet assigned

Serial No.: 10/607,577) Our Ref: B-5134 621037-8

Filed: June 25, 2003

For: "DYNAMIC CONTROL OF RESOURCE USAGE IN A MULTIMODAL SYSTEM") Date: February 10, 2004

CLAIM TO PRIORITY UNDER 35 U.S.C. 119

MAIL STOP MISSING PARTS
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

[X] Applicants hereby make a right of priority claim under 35 U.S.C. 119 for the benefit of the filing date(s) of the following corresponding foreign application(s):

COUNTRY	FILING DATE	SERIAL NUMBER
GREAT BRITAIN	28 June 2002	0215118.1

[ ] A certified copy of each of the above-noted patent applications was filed with the Parent Application No.______________.

[X] To support applicant's claim, a certified copy of the above-identified foreign patent application is enclosed herewith.

[ ] The priority document will be forwarded to the Patent Office when required or prior to issuance.

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first-class mail in an envelope addressed to the "Mail Stop Missing Parts, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450", on February 10, 2004

by Suzanne Johnston

Respectfully submitted,

Richard P. Berg
Attorney for Applicant
Reg. No. 28,145

LADAS & PARRY
5670 Wilshire Boulevard
Suite 2100
Los Angeles, CA 90036
Telephone: (323) 934-2300
Telefax: (323) 934-0202

USSN 10/607,577

The Patent Office

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed K Behen

Dated 16 June 2003

VSSN 10/607, 577

Patents Form 1/77

Patents Act 1977
(Rule

The Patent Office

THE PATENT OFFICE
M
28 JUN 2002
NEWPORT

1/77

01JUL02 E729849-1 D01463
P01/7700 0.00-0215118.1

The Patent Office

Cardiff Road
Newport
South Wales
NP10 8QQ

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

1. Your reference — 200206482-1 GB

2. Patent application number — 0215118.1 — 28 JUN 2002

3. ... or of each applicant *(underline all surnames)*

Hewlett-Packard Company
3000 Hanover Street
Palo Alto
CA 94304, USA

Patents ADP number *(if you know it)* — 49658800I

If the applicant is a corporate body, give the country/state of its incorporation — Delaware, USA

4. Title of the invention — Dynamic Resource Allocation in a Multimodal System

5. Name of your agent *(if you have one)*

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

Robert Squibbs
Hewlett-Packard Ltd. IP Section
Filton Road, Stoke Gifford
Bristol BS34 8QZ

Patents ADP number *(if you know it)* — 7928187001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*
   a) any applicant named in part 3 is not an inventor, or
   b) there is an inventor who is not named as an applicant, or
   c) any named applicant is a corporate body.
   See note (d))

Yes

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	7
Claim(s)	2
Abstract	1
Drawing(s)	4+4

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	Fee Sheet

11. I/We request the grant of a patent on the basis of this application.

Signature R F Griffiths   Date 28/6/2002

12. Name and daytime telephone number of person to contact in the United Kingdom   Tony Judd Tel: 0117-312-8026

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) *If you need help to fill in this form or you have any questions, please contact the Patent Office on 08459 500505.*

b) *Write your answers in capital letters using black ink or you may type them.*

c) *If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.*

d) *If you have answered 'Yes' Patents Form 7/77 will need to be filed.*

e) *Once you have filled in the form you must remember to sign and date it.*

f) *For details of the fee and ways to pay please contact the Patent Office.*

Dynamic Resource Allocation in a Multimodal System

Field of the Invention

The present invention relates to dynamic resource allocation in a multimodal system

Background of the Invention

Multimodal systems are systems which permit a user to provide input in different modalities, such as speech or gesture, in parallel, in sequence or as alternatives. The processing of an input modality is typically split up into a number of tasks carried out by corresponding functionality, herein referred to as task entities. The results of processing of input via one modality can be combined or fused with the results obtained from the processing of other modalities at any stage in the processing chain and is not restricted to being combined at the top level by the application to which the inputs are directed.

The processing demands for processing modalities such as speech can be very high if, for example, a large vocabulary is to be catered for and this has restricted the adoption of modalities such as speech as input interfaces for mobile devices which typically have very limited processing power and memory available. However, advances in wireless communication, ad hoc networks and human language technologies are set to enable mobile devices to offload processing tasks requiring specialized or powerful processing resources to infrastructure-based task entities. Figure 1 of the accompanying drawings illustrates a multimodal input system for a mobile device in which the symbolic recognition and syntactic analysis tasks involved in processing speech and gesture modalities are carried out by remote task entities. As can be seen, the outputs of the feature-extraction task entities of the mobile device are passed to the remote symbolic-recognition task entities over a communication channel; similarly, the outputs of the remote syntactic-analysis task entities are passed to the semantic-analysis task entities of the mobile device over the same or another communication channel. The setting up of the ad hoc organization of local and remote task entities is effected by a modality manager of the mobile device.

As a consequence, it may be expected that in the near future mobile device users will be able to use a plethora of interaction modalities such as speech, gesture recognition, etc. Users will also expect that their appliances will be able to to interact seamlessly, providing a multimodal user interface onto services and information regardless of the communication technology used by the device (for example, technologies such as 3G cellular, 802.11 wireless LAN, and Bluetooth).

In a world of disaggregated computing, the bandwidth between input clients (such as, but not limited to, mobile devices) and computing resources serving as task entities will dramatically influence where and to what degree multimodal input (with or without fusion) can be carried out effectively. At certain points in the communications infrastructure used by the input clients, bandwidth is likely to be less than needed. For example, where a mobile device has a collection of co-operating input clients that utilise internet-based task entities via an 802.11 network to process multiple input modalities, the bandwidth of the interconnection between the mobile device and the task entities will be influenced by other users in the local vicinity and the environment. A fall in the available bandwidth will impact all modalities currently being handled.

It is an object of the present invention to facilitate multimodal input in systems subject to resource restrictions.

Summary of the Invention

According to one aspect of the present invention, there is provided a method for dynamically allocating a resource used by task entities respectively involved in processing different input modalities, wherein the resource is dynamically allocated between the task entities according to one or more of the following:

- actual usage of the different modalities by a user;
- confidence in the results of processing of each of the modalities;
- pragmatic information on mode usage.

Pragmatic information on mode usage provides a measure of how the target application is set up to use input from different modes – in other words, whether input from one modality

is more important or useful than that from another modality, at least in the current application context.

The resource concerned is, for example, communication bandwidth or processing power.

The present invention also envisages systems for implementing the foregoing method.

Brief Description of the Drawings

Embodiments of the invention will now be described, by way of non-limiting example, with reference to the accompanying diagrammatic drawings, in which:

- **Figure 1** is a diagram of a mobile device with two input modalities where certain processing tasks in respect of those modalities are carried out on remote resources;
- **Figure 2** is a diagram illustrating the control of the allocation of communication bandwidth between task entities associated with different input modalities;
- **Figure 3** is a diagram similar to Figure 1 but showing bandwidth allocation control between modalities for first and second communication channels between the mobile device and the remote resources; and
- **Figure 4** is a diagram similar to Figure 3 but for the case of only a single communication channel existing between the mobile device and the remote resources.

Best Mode of Carrying Out the Invention

Figure 2 illustrates an embodiment of the present invention in which task entities have been organized by a modality manager to provide viable processing stacks for first and second input modalities. The processing stack of each input modality includes a respective pair of task entities that are linked via a communication channel common to both modalities. Bandwidth restrictions on the communication channel linking task entities of task-entity pair thus have the potential of affecting processing of both modalities.

However, in the Figure 2 arrangement a bandwidth moderator is provided to control the relative usage of the communication channel by the task entities of the two modalities. The bandwidth moderator receives inputs regarding input mode usage by the user, the modal requirements of the dialogue manager and application, and confidence in the recognition process for each modality. The first of these inputs can be derived from any processing stage up the processing stack formed by the task entities of each modality though generally the input will be derived at the stage controlled by the bandwidth moderator; the second input comes from the application and/or dialogue and/or pragmatic manager; and the third input can be an overall confidence measure from the application and/or dialogue and/or pragmatic manager top-level or a more local confidence measure from a task entity either controlled by the bandwidth moderator or a task entity receiving the output from an entity controlled by the bandwidth moderator. By way of example of a locally-derived third input, the syntactic-analysis task entity may monitor its own performance and if it is not confident that the correct sentence is represented in the word or phoneme lattice, then it indicates this to the associated bandwidth moderator with a view to getting increased bandwidth to represent sentences.

Whilst all three inputs are preferably provided to the bandwidth moderator, it is possible for the moderator to operate using just any two or any one of the inputs. Additional inputs may also be provided to the bandwidth moderator.

The bandwidth moderator uses the inputs it receives to determine the allocation of the channel bandwidth between the two modalities in order to seek to optimize overall input performance. For example:

- if a person is only using speech, when both speech and gesture recognition is available, then the bandwidth moderator allocates less bandwidth resource to gesture recognition;
- if speech recognition is found to be poor (a low confidence score is measured) increasing the data generated in the lower speech-modality task entities and allocating more bandwidth for passing on this data may well result in overall input performance gains outweighing any loss in gesture recognition capability resulting from the reduced data flow in the gesture modality processing stack.

In the present embodiment bandwidth allocation is effected by the moderator by controlling the amount of data output by the task entities that use the bandwidth-limited communication channel. How this is done depends on the type of task being carried out by each entity. For example, where the task entities concerned are sensors, the sampling rates of the sensors can be changed relative to each other to favour one modality over the other as required by the bandwidth moderator. If the task entities being controlled effect feature extraction then the bandwidth moderator can be arranged to control the number of features extracted for each modality. Similarly, if the task entities controlled by the bandwidth moderator effect syntactic and semantic analysis, then the depth and breath of the word or phoneme lattices can be controlled.

Whilst generally the task entities using the communications channel will be at the same level in the processing stacks of each modality, this is not necessarily the case as the moderator can be arranged to understand how to control different types of task entity to effect the desired bandwidth allocation. Furthermore, it will be appreciated that the bandwidth moderator can be arranged to allocate bandwidth between more than two modalities. Again, whilst the resource allocated by the moderator in the Figure 2 example is channel bandwidth, the moderator can be used to allocate other limited resources between modalities such as processing power and/or memory.

Figure 3 illustrates an arrangement in which both the feature-extraction task entities of two modalities share a first communication channel to respective symbol-recognition task entities, and the syntactic-analysis task entities of these modalities share a second communication channel to respective semantic-analysis task entities.

The allocation of the bandwidth of the first communication channel between the two feature-extraction task entities is controlled by a first bandwidth moderator whilst the allocation of the bandwidth of the second communication channel between the two syntactic-analysis task entities is controlled by a second bandwidth moderator. It would be possible simply to have the first and second bandwidth moderators work independently, each operating as described for the moderator of Figure 2. However, instead provision is

made for global coordination of the two moderators by a third, global, moderator. The role of the global moderator is to guide the first and second moderators in making their allocations. For example, the global moderator may determine that whilst the first moderator should favour the speech feature-extraction task entity over the gesture feature-extraction task entity, the second moderator should be more even-handed between the syntactic-analysis task entities of the two modalities. The first and second moderators make their final allocations taking into account respective local activity in the task entities they control; the first and second moderators may also take account of the allocations made by each other.

Of course, a single, global, moderator could be used to directly control allocation of bandwidth for both the first and second channels without the use of the local first and second moderators described above.

Instead of there being two separate communication channels at respective levels in the processing stacks of the two modalities, it may be that only a single channel is available both for communication between the feature-extraction task entities and the symbol-recognition task entities and for communication between the syntactic-analysis task entities and the semantic-analysis task entities. In this case, the general configuration of moderators shown in Figure 3 can still be employed with the global moderator now determining, for example, the allocation of bandwidth between the two processing-stack levels involved and the first and second moderators then each effecting a subordinate allocation at a respective one of these levels. An alternative arrangement of moderators is depicted in Figure 4 where the global moderator determines allocation between modalities and each modality has an associated moderator that effects a subordinate allocation between the two concerned levels of the processing stack handling the modality.

It will be appreciated that many variants are possible to the above described embodiments of the invention.

With regard to the location of the moderators themselves, these can be located locally or remote from the task entities they control.. However, at least notionally, the resource moderators can be considered as part of the modality manager of the device. It may be noted that a resource moderator can be arranged to restrict resource access to sero for a particular moderator in appropriate circumstances, thereby effectively eliminating that modality; preferably, however, the presence or absence of any particular modality is determined by higher-level functionality of the modality manager and the resource managers are arranged always to provide at least a minimum resource level to each modality that the higher-level functionality of the modality manager has decided should be present.

Whilst the particular task entity instances used in each modality processing stack can be predetermined or can be constituted by an ad hoc collection of available instances under the control of the modality manager, it is also possible to arrange for some or all of these entity instances to be predetermined (where all task entity instances are predetermined, the modality manager is not involved in organizing task entities to form viable modality processing stacks).

Although in the above described embodiments resource allocation is effected by controlling operation of the task entities to vary their usage of the resource, it will be appreciated that allocation can effected in other ways such as by limiting data delivery to the resource from the task entity subject of regulation either by queuing the data or by selective culling of that data.

CLAIMS

1. A method for dynamically allocating a resource used by task entities respectively involved in processing different input modalities, wherein the resource is dynamically allocated between the task entities according to one or more of the following:

- actual usage of the different modalities by a user;
- confidence in the results of processing of each of the modalities;
- pragmatic information on mode usage.

2. A method according to claim 1, wherein the resource is communication bandwidth.

3. A method according to claim 1, wherein the resource is processing power.

4. A method according to claim 1, wherein the resource is memory.

5. A method according to any one of the preceding claims applied to each of two separate resources each used by different respective entities of said different input modalities, the allocation of the two resources being independent of each other.

6. A method according to any one of claims 1 to 4 applied to each of two separate resources each used by different respective entities of said different input modalities, the allocation of the two resources being jointly controlled.

7. A method according to any one of claims 1 to 4 wherein said resource is used by multiple task entities for each modality, the resource being first allocated between modalities and then between task entities in the same modality.

8. A method according to any one of claims 1 to 4 wherein said resource is used by multiple task entities for each modality, the resource being first allocated between different groups of equivalent task entities of different modalities and then between task entities of the same group.

9. A method according to any one of the preceding claims, wherein resource allocation is effected by controlling operation of the task entities to vary their usage of the resource.

10. A bandwidth allocation method comprising the steps of:

(a) establish a connection between an adhoc collection of co-operating devices and a network based service or series of services,

(b) allot available bandwidth to each modality based on a prior knowledge of requirements and usage,

(c) monitor mode usage and confidence in allotted recognition task,

(d) monitor mode usage and confidence in recognition across all tasks,

(e) assess pragmatic and application usage of modes,

(f) collectively moderate bandwidth allocation between devices and network services to favour modes with high activity or poor performance.

ABSTRACT

Dynamic Resource Allocation in a Multimodal System

A limited resource, such as communication bandwidth or processing power, is dynamically allocated between task entities that are respectively involved in processing different input modalities. The allocation is effected in dependence on one or more of the actual usage of the different modalities by a user, the confidence in the results of processing of each of the modalities, and pragmatic information on mode usage.

(Figure 2)

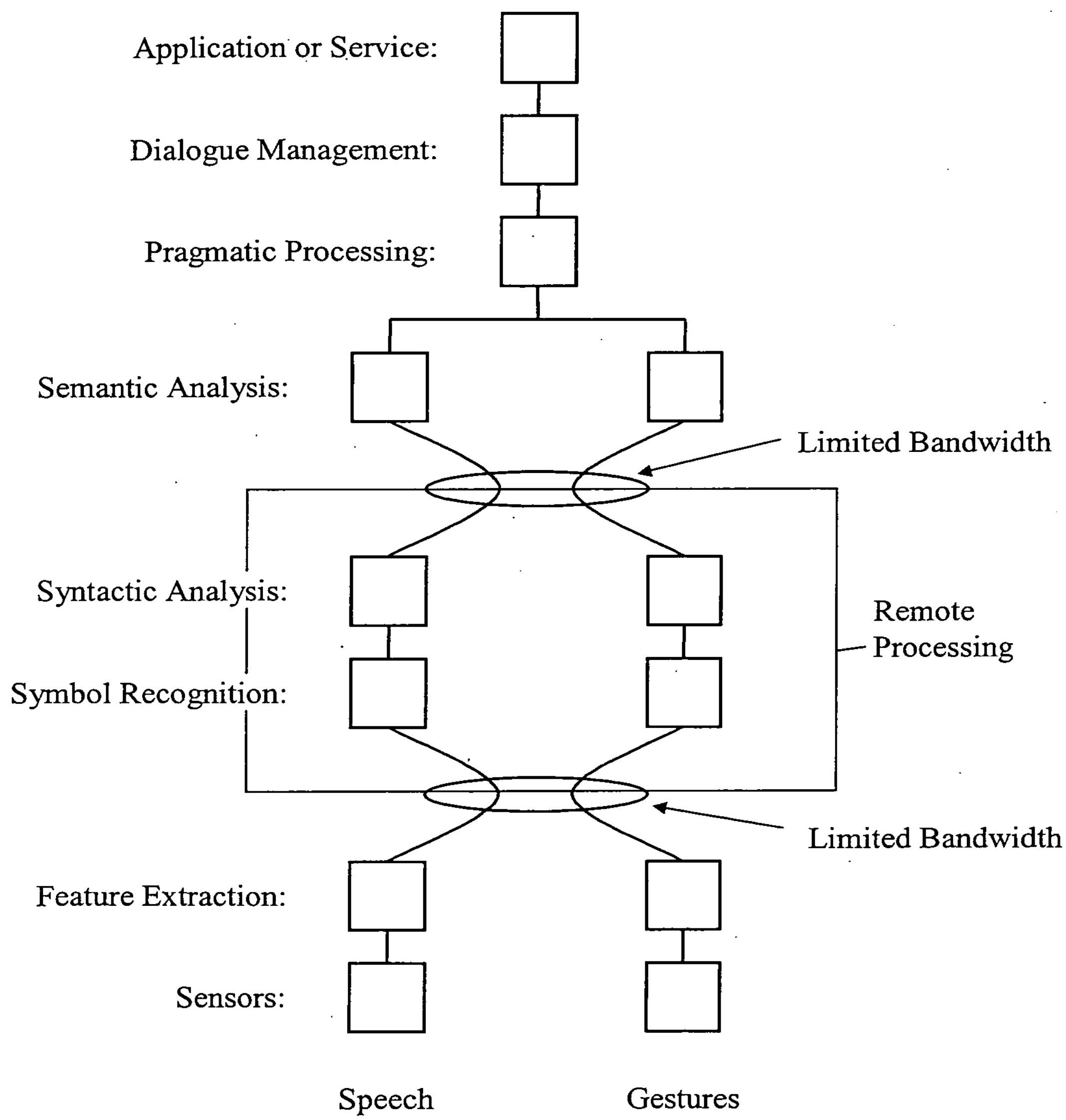
Application or Service:
Dialogue Management:
Pragmatic Processing:
Semantic Analysis:
Limited Bandwidth
Syntactic Analysis:
Symbol Recognition:
Remote Processing
Limited Bandwidth
Feature Extraction:
Sensors:
Speech
Gestures

Figure 1

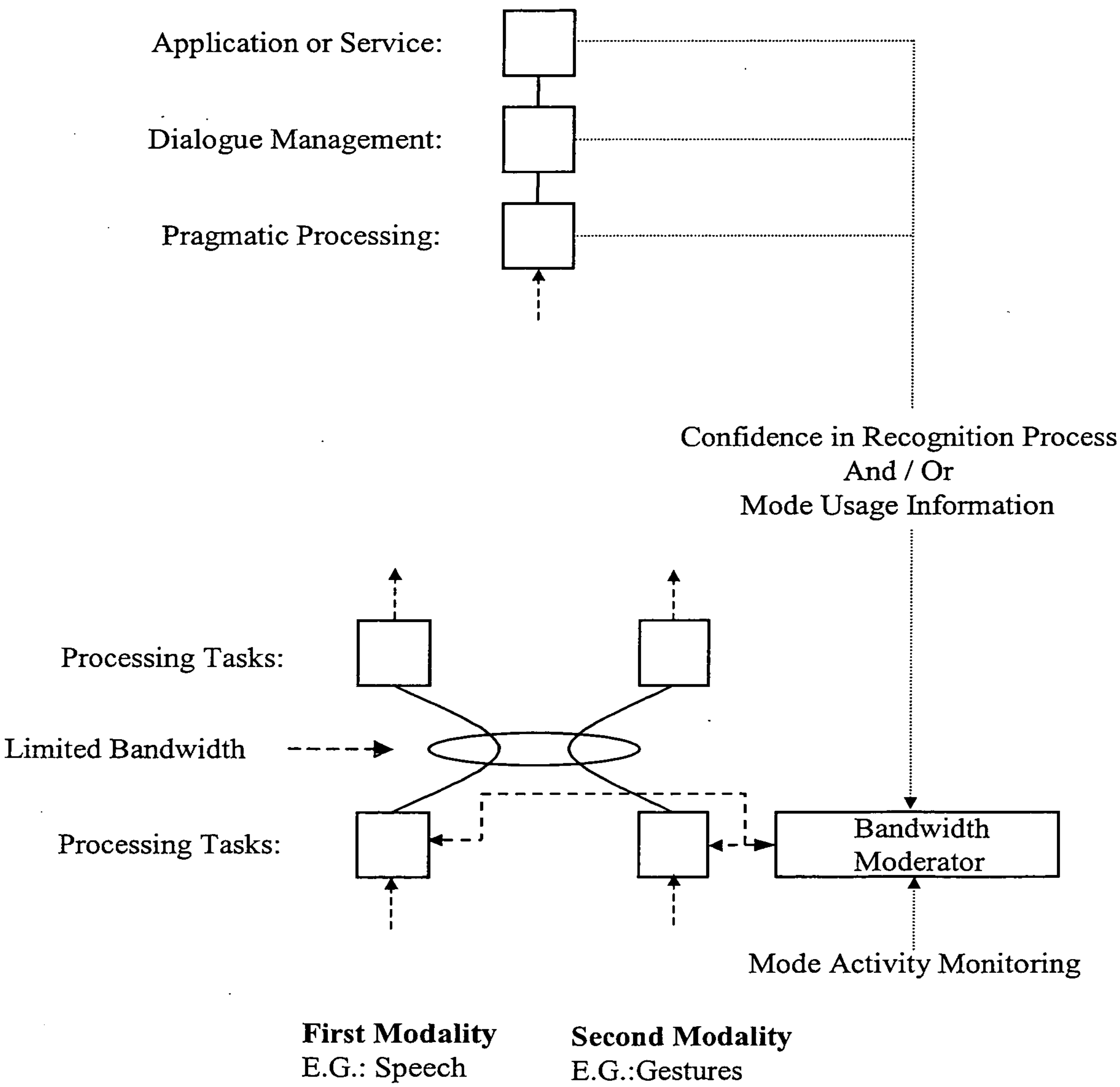
Application or Service:
Dialogue Management:
Pragmatic Processing:
Confidence in Recognition Process
And / Or
Mode Usage Information
Processing Tasks:
Limited Bandwidth
Processing Tasks:
Bandwidth
Moderator
Mode Activity Monitoring
First Modality
E.G.: Speech
Second Modality
E.G.:Gestures

Figure 2

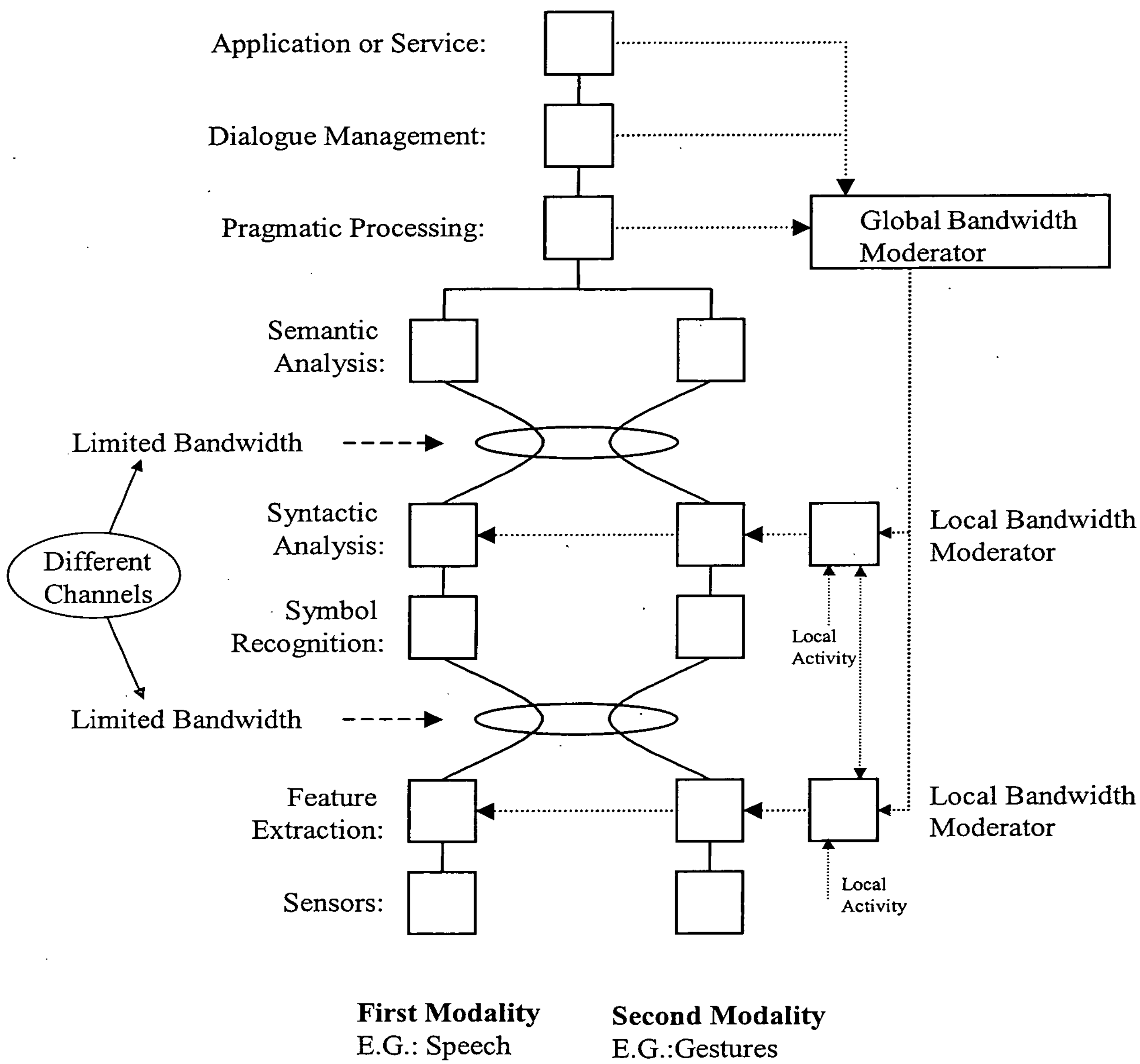
Application or Service:
Dialogue Management:
Pragmatic Processing:
Global Bandwidth Moderator
Semantic Analysis:
Limited Bandwidth
Syntactic Analysis:
Different Channels
Symbol Recognition:
Local Activity
Local Bandwidth Moderator
Limited Bandwidth
Feature Extraction:
Local Bandwidth Moderator
Sensors:
Local Activity
First Modality
E.G.: Speech
Second Modality
E.G.:Gestures

Figure 3

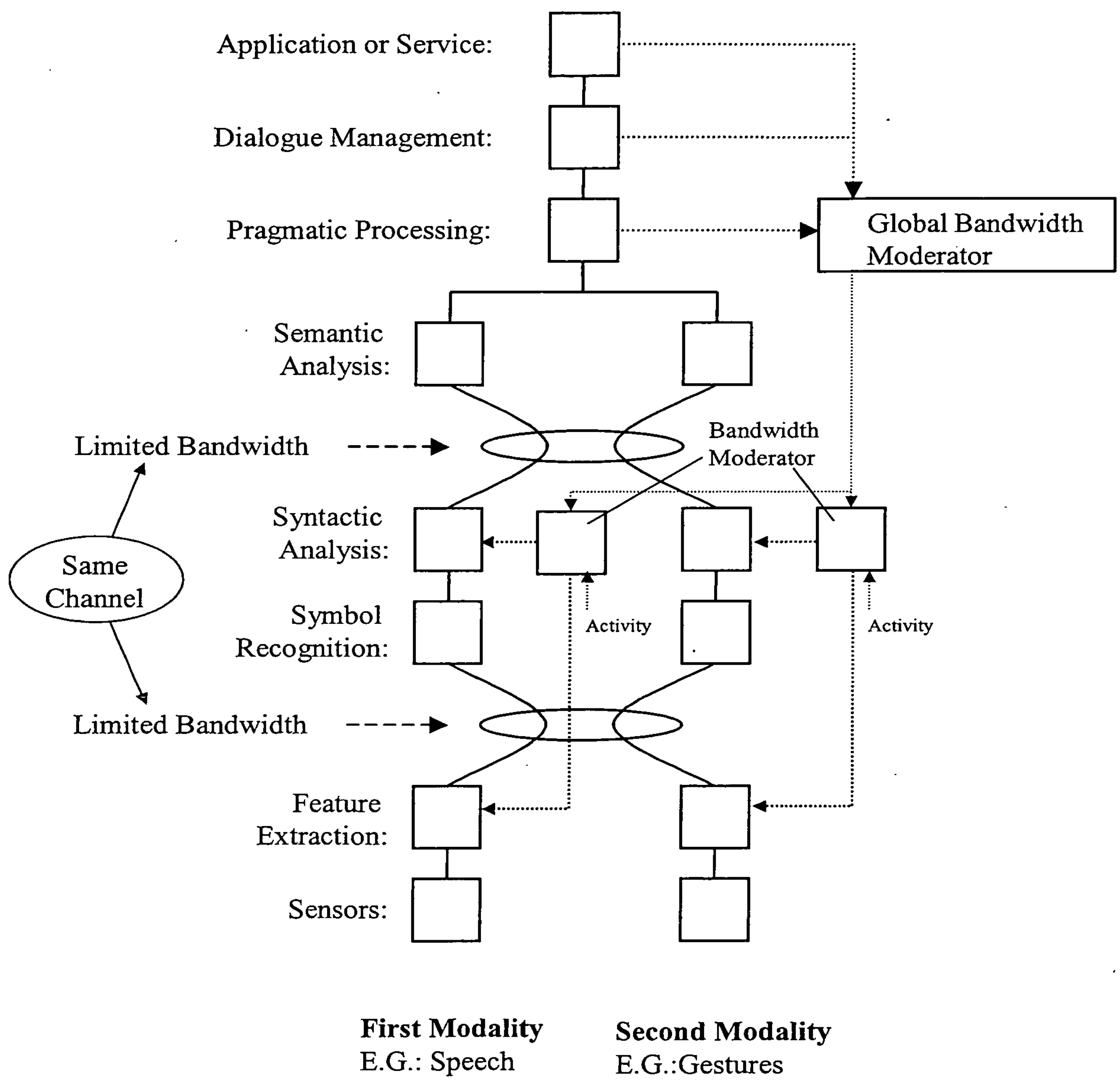
Application or Service:
Dialogue Management:
Pragmatic Processing:
Global Bandwidth Moderator
Semantic Analysis:
Limited Bandwidth
Bandwidth Moderator
Syntactic Analysis:
Same Channel
Symbol Recognition:
Activity
Activity
Limited Bandwidth
Feature Extraction:
Sensors:
First Modality
E.G.: Speech
Second Modality
E.G.:Gestures

Figure 4